Localização em Ambientes Internos Baseada em Aprendizado Supervisionado Utilizando Estações de Rádio FM

David Alan de O. Ferreira[1], **Waldir S. da Silva Júnior**[1], **Celso B. Carvalho**[1]

[1]Faculdade de Tecnologia - Universidade Federal do Amazonas (UFAM)
Av. General Rodrigo Octávio, 1200, Coroado I
CEP 69067-005 - Manaus - AM - Brazil

ferreirad08@gmail.com, {waldirjr, ccarvalho_}@ufam.edu.br

Abstract. *For the location of mobile devices on wireless networks, three or more fixed devices must be installed, whose transmitted signals are used as location parameters. The problem with this approach is the increase in energy and monetary costs. Thus, the objective of this work is to propose a method of location using FM radio stations with a view to low cost and high precision. The tests were carried out in a domestic environment with approximately 30 m^2 and 15 reference points. As a result of the tests, the proposed QA-PCA-kNN method stood out when using 6 characteristics of the FM signals, providing the location with a mean error of 0.0688 meters and a standard deviation of 0.2536 and presenting an accuracy of 86.80%.*

Resumo. *Para a localização de dispositivos móveis em redes sem fio, deve-se instalar três ou mais dispositivos fixos, cujos sinais transmitidos são utilizados como parâmetros de localização. O problema desta abordagem é o aumento dos custos energético e monetário. Assim, o objetivo deste trabalho é propor um método de localização utilizando estações de rádio FM com vistas ao baixo custo e alta acurácia. Foram realizados testes em ambiente doméstico com aproximadamente 30 m^2 e 15 pontos de referência. Como resultados dos testes, o método proposto QA-PCA-kNN destacou-se ao utilizar 6 características dos sinais FM, provendo a localização com erro médio de 0,0688 metros e desvio padrão de 0,2536 e, apresentando acurácia de 86,80%.*

1. Introdução

A localização de dispositivos em ambientes internos é um tema de grande interesse no âmbito das redes sem fio, visto que pode contribuir para o aprimoramento de serviços baseados em localização (SBL) em diversas áreas, como doméstica, industrial e hospitalar, entre outras [Al-Fuqaha et al. 2015]. Embora a localização fornecida pelo GPS (*Global Positioning System*) seja adequada para ambientes externos, este sistema utiliza sinais transmitidos por satélites, o que compromete a acurácia da localização quando os dispositivos encontram-se em ambientes fechados [Salim et al. 2014].

A localização em ambientes internos é atualmente alvo de diversos trabalhos, muitos destes apresentam propostas baseadas em redes Wi-Fi (padrões IEEE 802.11b/g/n), devido à ampla presença dessas redes em locais públicos e privados [Cai et al. 2015, Khullar and Dong 2017, Kim et al. 2018]. No entanto, a fim de se obter um erro aceitável, estas propostas utilizam como parâmetros de localização o posicionamento de no mínimo

três pontos de acesso (APs) e o *Received Signal Strength Indicator* (RSSI) destes APs [Le et al. 2014], resultando em um sistema de localização de alto custo energético e monetário.

Além disso, as redes Wi-Fi operam na faixa de frequência ISM (*Industrial, Sientific and Medical*). Esta faixa de frequência é aberta e, portanto, é amplamente utilizada [Li et al. 2009]. Com isso os sinais Wi-Fi são expostos a interferências de outras tecnologias de rádio, como Zigbee e Bluetooth, além de outros equipamentos eletrônicos, como telefone sem fio operando na faixa de 2,4 GHz e fornos de micro-ondas [Danbatta and Varol 2019, Rappaport 2002]. As interferências tornam o sistema de localização ainda mais vulnerável a falhas, pois o RSSI em um canal sem fio é previamente caracterizado por flutuações, nomeadamente: perda por percurso (*path loss*), sombreamento (*shadowing*) e multipercurso (*multipath*).

Aliado ao avanço das técnicas de *beamforming*, o advento dos padrões IEEE 802.11ac/ad/ax traz maior velocidade na transmissão de dados. O *beamforming* é empregado a múltiplas antenas disponíveis em um AP que identificam a posição relativa dos dispositivos móveis conectados, para então efetuar transmissões direcionadas, permitindo que os sinais sejam mais fortes em direções específicas. Promissoramente, autores têm procurado utilizar a técnica de *beamforming* nos sistemas de localização, visando reduzir as interferências nos sinais transmitidos e o número de APs necessários [Wen and Liang 2015]. Contudo, na prática, os padrões IEEE 802.11b/g/n ainda são amplamente utilizados nos mais diversos setores [Kapetanovic et al. 2020], considerando-se também que, é mais comum encontrar dispositivos que funcionam apenas com estes padrões, tais como o módulo ESP8266 NodeMCU.

Diante do exposto, este trabalho tem por objetivo propor um método de localização para ambientes internos utilizando canais de rádio FM com frequências na banda de 87,8 a 108 MHz, aproveitando a qualidade do sinal recebido e a alta disponibilidade em áreas urbanas. Adicionalmente, combinando abordagens estatísticas no tratamento dos dados para reconhecimento de informações adequadas a serem utilizadas por algoritmos de aprendizado supervisionado. A principal contribuição deste trabalho é oferecer robustez e baixo custo. Além de viabilizar a localização a partir de dispositivos embarcados sem instalação de complexa infraestrutura.

Este artigo está organizado nas seguintes seções: A Seção II apresenta trabalhos importantes e diretamente relacionados com o tema deste artigo. Na Seção III são apresentados os conceitos relevantes para realização dos estudos. A Seção IV apresenta o método e a proposta deste artigo. A Seção V expõe os resultados obtidos e a Seção VI apresenta a conclusão da pesquisa.

2. Trabalhos relacionados

Em [Popleteev et al. 2012], os autores investigaram a viabilidade de um sistema de localização baseado em sinais transmitidos por estações de rádio FM (*Frequency Modulation*), aproveitando-se da infraestrutura previamente existente e dos sintonizadores disponíveis em muitos dispositivos móveis. Foram avaliados os algoritmos k*-Nearest Neighbors* (kNN), *Support Vector Machine* (SVM) e *Gaussian Process* (GP) para classificação e regressão de vetores com valores médios de RSSI. Cada valor médio foi calculado a partir de 10 medições e normalizado na faixa de 0 a 1. No ambiente com área de 50 m

$\times$ 25 m, o melhor resultado foi obtido com a utilização do classificador kNN, utilizando o RSSI de 45 estações de rádio FM e estimando 52% das posições com erro médio de localização nulo. No ambiente com área de 12 m $\times$ 6 m, o kNN também apresentou o melhor resultado, utilizando 50 estações de rádio FM e estimando 40% das posições com erro médio de localização nulo. Em experimentos adicionais constatou-se que os receptores de FM têm um consumo energético de 2,6 a 5,5 vezes menor que os módulos Wi-Fi.

Em [Ferreira et al. 2020], os autores propuseram um método de localização em ambientes internos utilizando redes Wi-Fi. O objetivo foi melhorar a precisão de localização, que é comprometida pela instabilidade das medições de RSSI. O método emprega a análise de quartis (*Quartile Analysis* - QA) na representação dos dados e o algoritmo kNN. Em um ambiente com área de 3,5 m $\times$ 3,56 m, 100% das posições testadas foram identificadas com erro médio de localização nulo a partir de 4 APs, $k = 1$ e 10 medições de RSSI por AP para o cálculo dos quartis. O resultado alcançado com $k = 1$ mostra que o método é uma contribuição importante e promissora na área de localização.

Em [Salamah et al. 2016], utilizou-se a análise de componentes principais (*Principal Component Analysis* - PCA) para transformar o conjunto dos valores de RSSI em um conjunto com novas características, a fim de eliminar características com menor importância. O desempenho do método proposto foi testado utilizando os classificadores kNN, SVM, *Decision Tree* e *Random Forest*. Os experimentos foram conduzidos em um ambiente real com 45 pontos de referência (RPs) utilizando *smartphones* para a coleta dos valores de RSSI de 6 APs. Os resultados mostram que o kNN com $k = 2$ e utilizando as três primeiras características obteve o melhor desempenho nos experimentos dinâmicos, apresentando um erro médio de localização de 1,71 m com precisão de 60% e de 3,0 m com precisão de 79%. A localização foi calculada pela média ponderada das coordenadas (centróide) dos k vizinhos mais próximos. Este cálculo permite localizar objetos em posições coincidentes com os RPs e em posições aleatórias.

O trabalho de [Popleteev et al. 2012] contribui para uma abordagem de baixo custo, visto que os gastos com a instalação de dispositivos fixos são suprimidos. Os trabalhos de [Ferreira et al. 2020] e [Salamah et al. 2016] apresentam, respectivamente, a QA e a PCA como abordagens estatísticas para o tratamento de dados com intuito de alcançar melhores resultados. Estas abordagens estatísticas combinadas podem ajudar na caracterização do comportamento dos sinais de rádio FM em um cenário de localização.

3. Referencial teórico

3.1. Rádio FM

A Rádio FM é a modalidade de serviço de radiodifusão sonora por meio de modulação FM. A modulação FM é muito utilizada por codificar o sinal de áudio com alta fidelidade, menos ruído e sintonia em dispositivos móveis, como *tablets* e *smartphones*. Para cada estação (*STAtion* - STA) transmissora, é alocada uma frequência central (f_c) e uma largura de banda de 200 kHz situada no espectro de radiofrequências entre 87,8 MHz e 108 MHz (legislação local). Assim, uma STA pode operar na faixa de $f_c \pm 100$ kHz.

Neste trabalho foi utilizado o dongle RTL-SDR para a aquisição de dados. Este dispositivo tem um custo monetário de até 3 vezes menor que os módulos XBee, considerando o valor de mercado nacional. Além disso, é necessária a aquisição de apenas

um único dongle RTL-SDR. A Figura 1 ilustra este dispositivo, originalmente projetado para ser um receptor de TV digital (sistema ISDB-T). Este dispositivo é compatível com o desenvolvimento de rádio definido por software (*Software Defined Radio* - SDR), visto que pode receber sinais de radiofrequência na faixa de 48,25 MHz a 863,25 MHz.

Figura 1. Receptor RTL-SDR ISDB-T da Knup (KP-T2).

3.2. Análise de Quartis

A análise de quartis (QA) é um método estatístico utilizado para avaliar a tendência central, a dispersão e a posição das observações nos dados. Os quartis particionam um conjunto parcialmente ordenado (poset) em quatro partes iguais. O primeiro quartil ($Q_{\frac{1}{4}}$) ou quartil inferior delimita as 25% menores observações, o segundo quartil ($Q_{\frac{2}{4}}$) ou mediana separa as 50% menores das 50% maiores observações, e o terceiro quartil ($Q_{\frac{3}{4}}$) ou quartil superior delimita as 25% maiores observações [Joarder and Firozzaman 2001].

São encontradas na literatura diversas definições para o cálculo dos quartis [Langford 2006]. Dessa forma, uma equação generalizada para o cálculo computacional/estatístico dos quartis é definida neste artigo. A partir de um poset $(S, \leq)$, o valor do quartil Q_p é estimado através de regressão linear, entre os elementos $x_{\lfloor i \rfloor}$ e $x_{\lfloor i \rfloor+1}$, em que a posição $\lfloor i \rfloor$ é determinada em função da porcentagem de observações p, conforme a equação:

$$\begin{aligned} Q_p = x_{\lfloor i \rfloor} + (x_{\lfloor i \rfloor+1} - x_{\lfloor i \rfloor})(i - \lfloor i \rfloor) \\ com \quad i = (n-1)p + 1 \end{aligned} \tag{1}$$

onde n é o número de elementos em $(S, \leq)$ e $\lfloor i \rfloor$ é a parte inteira do índice i.

Para dados altamente assimétricos e/ou afetados por *outliers*, o $Q_{\frac{2}{4}}$ é mais eficiente que a média, e não necessita, previamente, de uma análise exploratória. Além disso, o intervalo interquartil ($IQR = Q_{\frac{3}{4}} - Q_{\frac{1}{4}}$) é uma medida estatística relativamente robusta frente ao desvio padrão [Mosteller and Tukey 1977].

3.3. Análise de Componentes Principais

A análise de componentes principais (PCA) foi introduzida por Karl Pearson em 1901 e, atualmente, é um método estatístico amplamente utilizado para analisar dados multivariados [Pearson 1901]. A PCA determina as direções que apresentam as maiores variações

dos dados em um espaço m-dimensional, em que m é o número de variáveis que descrevem o conjunto de dados. Estas direções ou componentes são dispostas em ordem decrescente de acordo com suas variações, onde a componente de maior variação possui maior importância. Com isso, pode-se eliminar as componentes de menor importância e minimizar a dimensão dos dados [Fang and Lin 2012].

De uma forma geral, a PCA transforma um conjunto de variáveis correlacionadas/redudantes em um novo conjunto de variáveis não-correlacionadas. Para o cálculo da PCA, as n observações de cada variável do conjunto de dados ($D_{n\times m}$) são centralizadas subtraindo-se de cada observação a média das respectivas variáveis e, em seguida, calcula-se a matriz de covariância ($C_{m\times m}$) entre estas variáveis. A covariância entre duas variáveis quaisquer x e y, já centralizadas, é definida como:

$$\sigma^2_{xy} = \frac{1}{n}\sum_{i=1}^{n}((x_i - \mu_x)(y_i - \mu_y)) \tag{2}$$

onde μ_x e μ_y são as médias das variáveis x e y, respectivamente.

Posteriormente, calcula-se os autovalores (λ) e os autovetores (v) da matriz $C_{m\times m}$. Os autovalores representam a variância de cada componente e os autovetores organizados em colunas ($V_{m\times m}$) representam as transformações lineares. Assim, o novo conjunto de variáveis ($P_{n\times m}$) é calculado a partir da combinação linear das m variáveis correlacionadas e dos coeficientes/elementos de cada autovetor:

$$P = DV \tag{3}$$

3.4. *k*-Nearest Neighbors

O k-*Nearest Neighbors* (kNN) é um dos algoritmos de classificação mais simples no âmbito da aprendizagem supervisionada [Cover and Hart 1967]. É não-paramétrico, ou seja, não necessita que os dados apresentem distribuição específica (ex.: gaussiana ou exponencial) e possui como hiperparâmetro o número de vizinhos mais próximos (k). Em geral, um grande número de vizinhos pode reduzir ou aumentar o desempenho do algoritmo, estando sujeito à extração de características discriminantes para representação dos dados. Ou seja, se houver muito ruído nos dados, um valor alto para o k tornará o classificador muito sensível ao ruído. Por outro lado, se os dados são eficientes, um valor alto para o k irá minimizar a sensibilidade ao pouco ruído.

No treinamento, o modelo de classificação é criado por um conjunto de instâncias (ou vetores de características) previamente classificadas e o valor de k é definido empiricamente para uma melhor precisão. Na classificação, uma instância de teste (nova instância) é introduzida no classificador treinado. Tradicionalmente, o classificador examina as classes das k instâncias de treino mais próximas (similares) à instância de teste, baseando-se em métricas de distância, como Manhattan, Euclidiana ou Minkowski. Posteriormente, atribui a instância de teste à classe majoritária, ou seja, à classe mais representada pelas k instâncias de treino [Lantz 2015].

3.5. Gaussian Naive Bayes

O *Gaussian Naive Bayes* (GNB) é um classificador probabilístico muito utilizado no aprendizado de máquina. Este algoritmo possui alta eficiência computacional e

fundamenta-se no teorema de Bayes desenvolvido por Thomas Bayes (1701-1761).

As predições são realizadas a partir das probabilidades *a posteriori* de uma instância de teste pertencer a cada classe e, em seguida, atribuindo a instância à classe de maior probabilidade [Gonzalez and Woods 2009]. A probabilidade $P(c_i|X)$ da classe c_i, dado uma instância de teste ou vetor de características $X = (x_1, x_2, ..., x_n)$, é definida como:

$$P(c_i|X) = \frac{P(X|c_i)P(c_i)}{P(X)} \tag{4}$$

Considerando que as probabilidades *a priori* de todas as classes são iguais, tem-se $P(c_i) = \frac{1}{m}$, em que m é o número de classes; e dado que a probabilidade $P(X)$ está associada apenas a instância de teste, a função de decisão pode ser reduzida e representada como:

$$\hat{c} = argmax_{c_i} \prod_{j=1}^{n} P(x_j|c_i) \tag{5}$$

onde n é o número de características contidas no vetor X, j é o índice da j-ésima característica x_j, e a classe c_i que maximiza esta função corresponde a classe estimada $\hat{c}$.

Pressupondo que os dados apresentem distribuição gaussiana, utiliza-se a função de densidade de probabilidade (*Probability Density Function* - PDF) da distribuição normal para estimar $P(x_j|c_i)$:

$$P(x_j|c_i) = \frac{1}{\sigma_{c_i}\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_j - \mu_{c_i}}{\sigma_{c_i}}\right)^2} \tag{6}$$

À medida que as características extraídas tendem a descrever uma distribuição mais próxima da normal, e que não existe uma correlação entre elas, melhor é o desempenho do algoritmo [Brownlee 2016].

4. Material e métodos

4.1. Cenário experimental e aquisição de dados

Para o desenvolvimento deste trabalho, foram conduzidos experimentos em um ambiente interno com dimensões de 6,10 m $\times$ 4,80 m, contendo 15 RPs, conforme ilustrado na Figura 2. Ao invés das medições de RSSI, o dongle RTL-SDR disponibiliza o ganho de potência (G_p) de cada STA. No entanto o G_p corresponde a uma relação entre a potência recebida (P_{rx}) e a potência de transmissão (P_{tx}), e é difinido como:

$$G_p[dB] = 10 \log_{10} \frac{P_{rx}[W]}{P_{tx}[W]} \tag{7}$$

Dispondo do espectro eletromagnético disponível, arbitrariamente, foram selecionadas 8 STAs para os experimentos. O dongle foi realocado sequencialmente em todos os RPs, e permaneceu imóvel ao longo de 2 minutos em cada RP. Foram coletadas 2000

medições do G_p de cada STA durante o tempo de imobilidade em cada RP. Após a coleta em todos os 15 RPs, os dados foram mesclados em um único arquivo de texto CSV, resultando em uma tabela (conjunto de dados) com 30000 linhas e 8 colunas. Sendo 2000 linhas por RP.

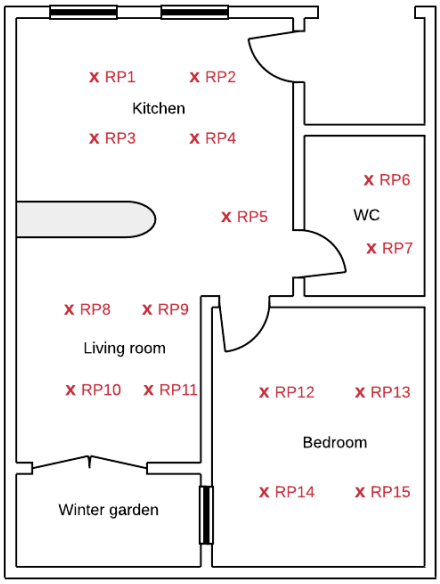

Figura 2. Cenário experimental com coordenadas dos pontos de referência.

4.2. Pré-processamento e transformação de dados

O principal objetivo desta fase é melhorar a qualidade dos dados coletados, visto que os dados podem apresentar problemas como grande quantidade de dados ausentes (*missing data*) e valores discrepantes (*outliers*), possivelmente decorrentes das interferências no sinal de radiofrequência. Adicionalmente, buscou-se extrair características e reduzir a dimensionalidade dos dados a fim de obter uma representação com características/atributos discriminantes.

Analisando o conjunto de dados, constatou-se que 8,6% do total de observações são dados ausentes. Entende-se que essas perdas esporádicas do sinal são interferências de fatores externos, como características da antena de recepção e obstáculos em uma determinada posição. Então, realizou-se uma imputação múltipla por equações encadeadas (*Multiple Imputation by Chained Equations* - MICE) considerando o mecanismo completamente aleatório dos dados ausentes (*Missing Completely at Random* - MCAR) [Azur et al. 2011, Harrell 2001]. Após a obtenção de um conjunto de dados completo, as observações dos atributos foram normalizadas utilizando a fórmula Min-Max:

$$x_{normalizado} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{8}$$

onde x é uma observação a ser normalizada e x_{min} e x_{max} são os valores mínimos e máximos dentre as observações do atributo, respectivamente. Esta fórmula corresponde a uma transformação linear do atributo em uma escala de 0 a 1 e mantém a distribuição das observações.

No tratamento de *outliers* e extração de características/atributos, criou-se um conjunto de instâncias (vetores de características) utilizando a QA. A cada 10 linhas correspondentes a um RP do conjunto de dados coletados, calculou-se os quartis de cada coluna, ou seja, uma instância baseia-se em 10 observações do G_p de cada STA. Uma instância é associada a um único RP e é representada por um vetor X de 24 características, ou seja, 3 características para cada STA:

$$X = (\overbrace{Q_{\frac{1}{4}}, Q_{\frac{2}{4}}, Q_{\frac{3}{4}}}^{STA_1}, \overbrace{Q_{\frac{1}{4}}, Q_{\frac{2}{4}}, Q_{\frac{3}{4}}}^{STA_2}, ..., \overbrace{Q_{\frac{1}{4}}, Q_{\frac{2}{4}}, Q_{\frac{3}{4}}}^{STA_8}) \tag{9}$$

Como o conjunto de dados coletados possui 2000 observações de cada STA em cada RP, foram criadas 200 instâncias para cada um dos 15 RPs. Esta abordagem elimina os *outliers* e extrai características que denotam o comportamento das medições brutas.

Em seguida, utilizou-se a PCA para garantir características relevantes e não correlacionadas/redundantes. O algoritmo da PCA, utilizado para transformação dos dados, retorna novas características que seguem uma ordem de importância. Esta ordenação viabilizou a redução de dimensionalidade dos dados, onde selecionou-se as 8 características mais relevantes.

Nos experimentos utilizou-se 75% das instâncias para treinamento e 25% para testes, visando impedir qualquer influência (viés) nos resultados, pois os dados de teste não devem ser utilizados no treino. Com intuito de obter um conjunto de treinamento balanceado, isto é, com o mesmo número de instâncias entre as classes, foi realizada uma amostragem aleatória estratificada (*stratified random sampling*).

4.3. Seleção do modelo de classificação

Para estimar as coordenadas associadas à instância de teste, ou seja, identificar a localização propriamente dita, foram considerados e avaliados os algoritmos de aprendizado supervisionado kNN e GNB utilizando o software MATLAB®. Em ambos os modelos, as coordenadas $(\hat{x}_i, \hat{y}_i)$ são estimadas a partir do centróide:

$$(\hat{x}_i, \hat{y}_i) = \left(\frac{\sum_{i=1}^{k} w_i x_i}{\sum_{i=1}^{k} w_i}, \frac{\sum_{i=1}^{k} w_i y_i}{\sum_{i=1}^{k} w_i} \right) \tag{10}$$

onde k é o número de RPs selecionados, (x_i, y_i) são as coordenadas dos RPs e w_i é o peso de cada RP. O cálculo do centróide permite localizar objetos em posições aleatórias dentro do ambiente. Nota-se que, neste trabalho busca-se estimar as coordenadas de posições coincidentes com os RPs, e considera-se que estas posições também podem ser testadas como quaisquer outras posições aleatórias, assumindo que quaisquer posições testadas são equiprováveis.

O kNN estima as coordenadas a partir dos RPs associados aos $k = 4$ vizinhos mais próximos, ponderados pelo inverso das distâncias Euclidianas calculadas $w_i = \frac{1}{d_i}$; e o GNB utiliza os $k = 4$ RPs com maiores probabilidades, ponderados pelas probabilidades calculadas $w_i = P(c_i|X)$. Estes modelos de classificação foram testados de forma empírica e o erro médio (EM) de localização foi utilizado como métrica de desempenho, a fim de medir a distância média entre as coordenadas reais e as coordenadas estimadas. O EM é dado em metros e definido como:

$$EM = \frac{1}{N}\sum_{i=1}^{N}\sqrt{(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2} \tag{11}$$

onde N é o número total de testes realizados.

A fim de obter uma conformidade entre erro de localização e baixo custo computacional, buscou-se determinar o número ótimo de características. Para tanto, o número de características utilizadas variou de 3 a 8. O desempenho do método proposto é comparado com a simples abordagem de médias no pré-processamento, denotada pela letra maiúscula M. Nesta abordagem, as médias de 10 medições são normalizadas pela fórmula Min-Max como em [Popleteev et al. 2012, Moghtadaiee and Dempster 2014]. O uso da média para reduzir as variações da medições reduzem o erro quadrático da medições, mas não necessariamente extraem características discriminantes.

5. Resultados e discussão

Inicialmente, o desempenho dos métodos foi analisado observando-se a influência do número de características no EM de localização. O número de características, necessário para minimizar o EM, está relacionado diretamente com o conhecimento adquirido para discriminação dos RPs. Os resultados de desempenho, em função do número de características, são apresentados na Figura 3.

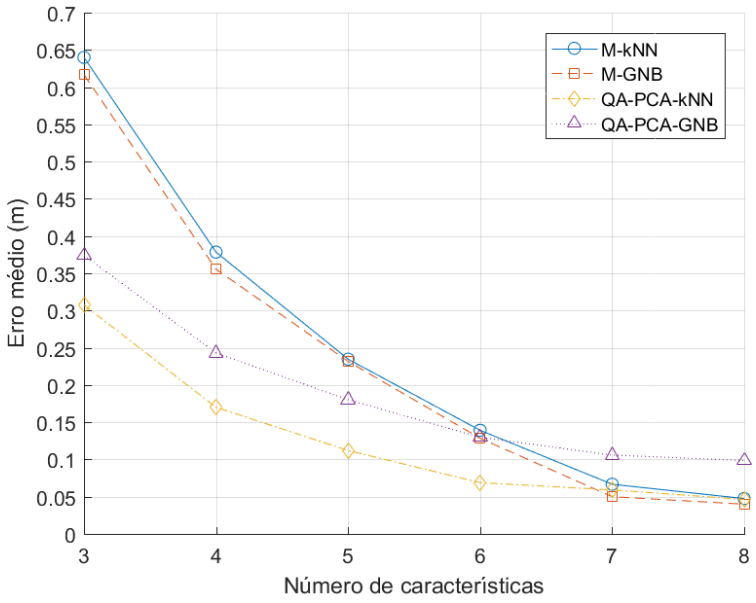

Figura 3. Influência do número de características no erro médio.

Nota-se nas curvas de avaliação de desempenho que houve maior redução no valor de EM quando a quantidade de características varia de 3 a 6. Utilizando de 3 e 6 características, o menor valor do EM foi alcançado pelo método QA-PCA-kNN com 0,0688 m, sendo superado apenas pelo método M-GNB com EM de 0,0399 m, utilizando 7 e 8 características. De 3 a 5 características, os métodos M-GNB e M-kNN obtiveram os

piores desempenhos com EM de até 0,6168 e 0,6400 m, respectivamente. O método QA-PCA-GNB obteve o segundo melhor desempenho de 3 a 5 características, mas obteve o pior desempenho com 7 e 8 características com EM de 0,1058 m.

Em seguida, buscou-se inferir a confiabilidade dos métodos testados a partir do desvio padrão das estimativas. As Tabelas 1, 2 e 3 apresentam as estatísticas das estimativas de localização utilizando 3, 6 e 8 características, respectivamente.

Tabela 1. Indicadores de desempenho para 3 características.

	mean (m)	std (m)	min (m)	max (m)
M-kNN	0,6400	0,8900	0	4,1060
M-GNB	0,6168	0,7611	0	3,8056
QA-PCA-kNN	0,3074	0,5518	0	3,7802
QA-PCA-GNB	0,3741	0,5413	0	3,7373

Tabela 2. Indicadores de desempenho para 6 características.

	mean (m)	std (m)	min (m)	max (m)
M-kNN	0,1391	0,4042	0	3,3879
M-GNB	0,1284	0,3553	0	2,8976
QA-PCA-kNN	0,0688	0,2536	0	3,2451
QA-PCA-GNB	0,1297	0,3232	0	3,2357

Tabela 3. Indicadores de desempenho para 8 características.

	mean (m)	std (m)	min (m)	max (m)
M-kNN	0,0474	0,2244	0	2,5022
M-GNB	0,0399	0,2057	0	3,2112
QA-PCA-kNN	0,0463	0,2069	0	3,2451
QA-PCA-GNB	0,0986	0,3146	0	3,2297

Observa-se que à medida que número de características aumenta, todos os métodos mostram-se mais robustos, com os desvios padrão cada vez menores. Nos testes com 3 e 6 características, destaca-se que os métodos baseados nos quartis e nas componentes principais (QA-PCA-kNN e QA-PCA-GNB) expressam os menores desvios padrão. Nos testes com 8 características, o método M-GNB obteve as melhores estatísticas.

Em fim, buscou-se a proporção de estimativas com erros nulos (0 m), ou simplesmente acurácia. As Figuras 4a e 4b apresentam a CDF (*Cumulative Distribution Function*) dos erros para vetores de 3 e 6 características, respectivamente.

Nestas CDFs, os métodos baseados no algoritmo k-Nearest Neighbors apresentaram as maiores acurácias. O QA-PCA-kNN alcançou a estimativa de 86,80% das posições com erro de localização nulo, enquanto o M-kNN estimou 80,27% das posições. A partir de um vetor de tamanho 8, o M-kNN com 91,87% superou a acurácia do QA-PCA-kNN de 90,80%, conforme ilustrado na Figura 4c. No entanto, a baixa dimensão dos vetores de

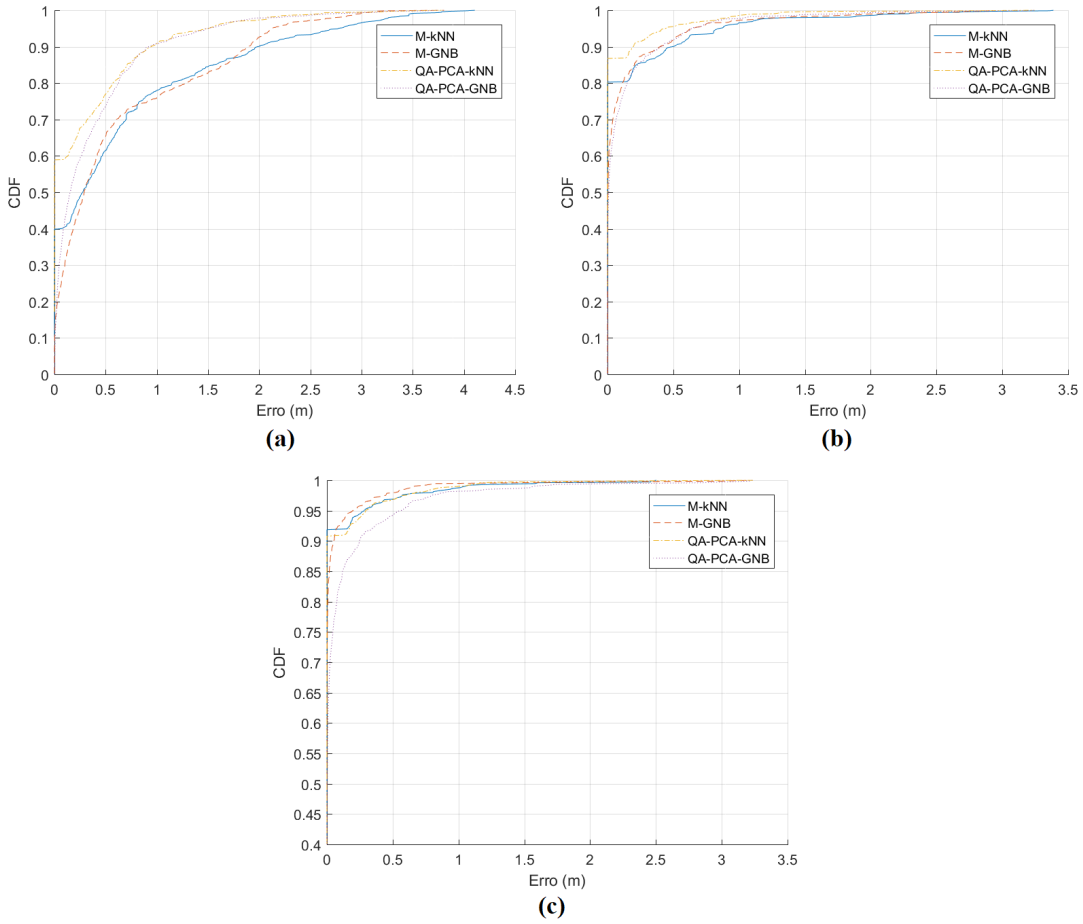

Figura 4. CDF (*Cumulative Distribution Function*) dos erros para (a) 3, (b) 6 e (c) 8 características.

características viabiliza ainda mais a utilização de sistemas embarcados, com menor capacidade de memória e baixo poder de processamento, como dispositivos de localização.

6. Conclusões

Neste trabalho foi proposto um método de localização em ambientes internos utilizando canais de rádio FM. Para isso, foram combinadas duas diferentes abordagens estatísticas na representação dos dados: Análise de Quartis (QA) e Análise de Componentes Principais (PCA). E avaliadas duas diferentes abordagens de aprendizado supervisionado para a tarefa de classificação: k-*Nearest Neighbors* (kNN) e *Gaussian Naive Bayes* (GNB).

A partir dos resultados obtidos nos experimentos, pode-se validar o método QA-PCA-kNN para estimar a localização utilizando o centróide, uma vez que destacou-se com vetores de até 6 características. Além de apresentar um erro médio de localização mínimo satisfatório de 0,0688 metros com desvio padrão de 0,2536 para a base de dados.

Agradecimentos

Esta pesquisa foi financiada pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) e pela Fundação de Amparo à Pesquisa do Estado do Amazonas (FAPEAM/ProgramaPPP); iv) fundo setorial de infraestrutura (CT-INFRA); v) MCT/CNPQ.

Referências

Al-Fuqaha, A., Guizani, M., Mohammadi, M., Aledhari, M., and Ayyash, M. (2015). Internet of things: A survey on enabling technologies, protocols, and applications. *IEEE Communications Surveys Tutorials*, 17(4):2347–2376.

Azur, M., Stuart, E., Frangakis, C., and Leaf, P. (2011). Multiple imputation by chained equations: What is it and how does it work? *International journal of methods in psychiatric research*, 20:40–9.

Brownlee, J. (2016). *Master Machine Learning Algorithms: Discover How They Work and Implement Them From Scratch*. Jason Brownlee.

Cai, X., Li, X., Yuan, R., and Hei, Y. (2015). Identification and mitigation of nlos based on channel state information for indoor wifi localization. In *2015 International Conference on Wireless Communications Signal Processing (WCSP)*, pages 1–5.

Cover, T. and Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27.

Danbatta, S. J. and Varol, A. (2019). Comparison of zigbee, z-wave, wi-fi, and bluetooth wireless technologies used in home automation. In *2019 7th International Symposium on Digital Forensics and Security (ISDFS)*, pages 1–5.

Fang, S. and Lin, T. (2012). Principal component localization in indoor wlan environments. *IEEE Transactions on Mobile Computing*, 11(1):100–110.

Ferreira, D., Souza, R., and Carvalho, C. (2020). Qa-knn: Indoor localization based on quartile analysis and the knn classifier for wireless networks. *Sensors*, 20(17):4714.

Gonzalez, R. and Woods, R. (2009). *Processamento Digital De Imagens*. ADDISON WESLEY BRA.

Harrell, F. E. (2001). *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. Springer New York, New York, NY.

Joarder, A. and Firozzaman, M. (2001). Quartiles for discrete data. *Teaching Statistics*, 23:86–89.

Kapetanovic, Z., Moore, G. E., Garman, S., and Smith, J. R. (2020). Classifying wlan packets from the rf envelope: Towards more efficient wireless network performance. In *Proceedings of the 4th International Workshop on Embedded and Mobile Deep Learning*, EMDL'20, page 13–18, New York, NY, USA. Association for Computing Machinery.

Khullar, R. and Dong, Z. (2017). Indoor localization framework with wifi fingerprinting. In *2017 26th Wireless and Optical Communication Conference (WOCC)*, pages 1–6.

Kim, K. S., Wang, R., Zhong, Z., Tan, Z., Song, H., Cha, J., and Lee, S. (2018). Large-scale location-aware services in access: Hierarchical building/floor classification and location estimation using wi-fi fingerprinting based on deep neural networks. *Fiber and Integrated Optics*, 37(5):277–289.

Langford, E. (2006). Quartiles in elementary statistics. *Journal of Statistics Education*, 14.

Lantz, B. (2015). *Machine Learning with R*. Packt Publishing, 2nd edition.

Le, W., Wang, Z., Wang, J., Zhao, G., and Miao, H. (2014). A novel wifi indoor positioning method based on genetic algorithm and twin support vector regression. In *The 26th Chinese Control and Decision Conference (2014 CCDC)*, pages 4859–4862.

Li, H., Syed, M., Yao, Y.-D., and Kamakaris, T. (2009). Spectrum sharing in an ism band: Outage performance of a hybrid ds/fh spread spectrum system with beamforming. *EURASIP J. Adv. Sig. Proc.*, 2009.

Moghtadaiee, V. and Dempster, A. (2014). Indoor location fingerprinting using fm radio signals. *Broadcasting, IEEE Transactions on*, 60:336–346.

Mosteller, F. and Tukey, J. W. (1977). *Data Analysis and Regression: a Second Course in Statistics*. pub-AW, pub-AW:adr.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2:559–572.

Popleteev, A., Osmani, V., and Mayora, O. (2012). Investigation of indoor localization with ambient fm radio stations. In *2012 IEEE International Conference on Pervasive Computing and Communications*, pages 171–179.

Rappaport, T. (2002). *Wireless communications: Principles and practice*. Prentice Hall communications engineering and emerging technologies series. Prentice Hall, 2nd edition. Includes bibliographical references and index.

Salamah, A. H., Tamazin, M., Sharkas, M. A., and Khedr, M. (2016). An enhanced wifi indoor localization system based on machine learning. In *2016 International Conference on Indoor Positioning and Indoor Navigation (IPIN)*, pages 1–8.

Salim, F., Williams, M., Sony, N., Dela Pena, M., Petrov, Y., Saad, A. A., and Wu, B. (2014). Visualization of wireless sensor networks using zigbee's received signal strength indicator (rssi) for indoor localization and tracking. In *2014 IEEE International Conference on Pervasive Computing and Communication Workshops (PERCOM WORKSHOPS)*, pages 575–580.

Wen, F. and Liang, C. (2015). Fine-grained indoor localization using single access point with multiple antennas. *IEEE Sensors Journal*, 15(3):1538–1544.